Trabalho 5: Detecção de Spam com Aprendizagem de Máquina

Introdução

O objetivo deste trabalho é utilizar a ferramenta Scikit-learn para executar dois algoritmos de aprendizagem de máquina em problema de classificação. Os algoritmos a serem comparados são: geração de árvores de decisão (*decision trees*) e bayes ingênuo (*naive bayes*). Você deve executar as partes 1 e 2 e preparar um relatório que deve ser entregue pelo Moodle. O resultado deste trabalho é portanto um relatório em formato PDF detalhando suas atividades, além de dois scripts em Python, conforme descrito nas partes 1 e 2 abaixo.

Parte 1 - Árvores de Decisão

Nesta parte, o objetivo é aprender uma árvore de decisão sobre o conjunto de dado Balance Scale. Para isso, você irá replicar os passos do tutorial disponível em http://dataaspirant.com/2017/02/01/decision-tree-algorithm-python-with-scikit-learn/. Replique todos os comandos apresentados nesse tutorial, até gerar a medida de acurácia (*accuracy*) sobre o conjunto de dados de treinamento. Defina todos os comandos para isso em um único script Python denominado balance_scale.py.

Parte 2 - Bayes Ingênuo (Naive Bayes)

Nesta parte, o objetivo é aprender um modelo bayesiano simples multinomial (Multinomial Naive Bayes) sobre o conjunto de dados 20-newsgroups. Para isso, você irá replicar os passos descritos no tutorial disponível em https://jakevdp.github.io/PythonDataScienceHandbook/05.05-naive-bayes.html. Replique todos os comandos apresentados nesse tutorial, até gerar a medida de acurácia (*accuracy*) sobre o conjunto de dados de treinamento. Para cada passo, descreva em seu relatório o objetivo dele. Defina todos os comandos para geração do modelo e dos testes em um único script Python denominado newsgroups.py.